Усереднені за 50 розбиттями показники наведено в табл. 1 та на рис. 1. Найвищу точність отримано для логістичної регресії (0,983), далі йдуть наївний Байєс (0,972) і метод k найближчих сусідів зі стандартизацією (0,964); дерево рішень глибиною 3 дало 0,905. Різниця між першими трьома методами (до 1,9 в. п.) не перевищує розкиду оцінок між розбиттями (стандартне відхилення 0,016-0,031), тому на цих даних не можна стверджувати переваги одного з них.

Таблиця 1 - Результати перехресної перевірки (середнє ± стандартне відхилення, 50 розбиттів)

Метод	Accuracy	Macro-F1
kNN (k=5), стандартизація	0,964 ± 0,031	0,964 ± 0,031
kNN (k=5), без стандартизації	0,711 ± 0,071	0,690 ± 0,074
Наївний Байєс (Гауса)	0,972 ± 0,032	0,973 ± 0,031
Логістична регресія	0,983 ± 0,016	0,983 ± 0,016
Дерево рішень (глибина 3)	0,905 ± 0,051	0,906 ± 0,051

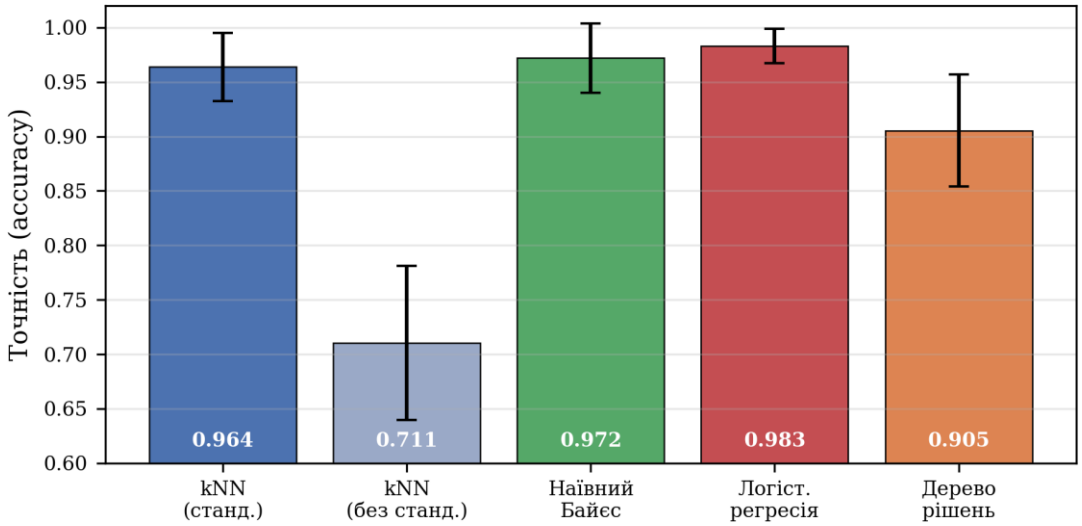


Рис. 1. Точність (accuracy) методів класифікації; позначено середнє та стандартне відхилення

Найбільш показовим є вплив масштабування: без стандартизації точність методу k найближчих сусідів становить 0,711, що на 25,3 в. п. нижче, ніж зі стандартизацією (макро-F1 – 0,690 проти 0,964). Причина полягає в тому, що в наборі ознаки мають різні порядки величин (зокрема, пролін вимірюється сотнями одиниць, тоді як інші показники – одиницями), і відстань (2) визначається переважно однією ознакою. Дерево рішень глибиною 3 поступається лінійним та ймовірнісним моделям, імовірно через обмежену глибину; цю гіпотезу в роботі не перевіряли.

Щоб перевірити, чи не залежить висновок про метод k найближчих сусідів від вибору k = 5, оцінку повторено за того самого протоколу для k від 1 до 21 (табл. 2). Точність зростає від 0,954 (k = 1) до 0,970 (k = 21), тобто в межах 1,6 в. п., що менше за стандартне відхилення оцінок за окремими розбиттями. Отже, на цих даних метод стійкий до вибору k, а головним чинником якості є стандартизація ознак.

Таблиця 2 – Залежність точності kNN (зі стандартизацією) від числа сусідів k

k	1	3	5	7	11	15	21
Accuracy	0,954	0,955	0,964	0,967	0,964	0,969	0,970
Станд. відхил.	0,035	0,036	0,031	0,028	0,027	0,029	0,025

Для логістичної регресії підсумкову матрицю помилок (табл. 3) отримано об'єднанням прогнозів на тестових блоках 10 незалежних 5-блокових розбиттів (усього 1780 прогнозів). Правильно класифіковано 1742 зі 1780 зразків (0,979). Усі помилки пов'язані з класом 2, який є найчисленнішим і за значеннями показників лежить між двома іншими: 27 його зразків віднесено до інших класів (до класу 3 – 21, до класу 1 – 6), ще 11 зразків інших класів помилково віднесено до класу 2. Зразки класів 1 і 3 між собою практично не змішуються. Це узгоджується з очікуванням, що найскладніше розділяти сусідні за хімічним складом групи.

Таблиця 3 - Матриця помилок логістичної регресії (сума за 10 розбиттями, n = 1780)

Фактичний клас \ прогноз	Клас 1	Клас 2	Клас 3
Клас 1	589	1	0
Клас 2	6	683	21
Клас 3	0	10	470

Результати показують, що для табличних лабораторних даних невеликого обсягу достатньо найпростіших моделей, а стандартизація ознак є обов'язковим кроком для методів, що ґрунтуються на відстанях. Запропонований протокол оцінювання (повторна стратифікована перехресна перевірка) можна застосувати й до даних про органолептичну якість вина [1]. Обмеження: невелика вибірка, добре розділені класи, відсутність налаштування гіперпараметрів і статистичних тестів між моделями; результати не переносяться автоматично на прогнозування експертної оцінки якості.

Висновки. 1. На наборі Wine логістична регресія, наївний Байєс і k найближчих сусідів зі стандартизацією досягли точності 0,96–0,98 без статистично підтвердженої різниці між ними. 2. Відмова від стандартизації знижує точність методу k найближчих сусідів приблизно на 25,3 в. п. 3. Результат kNN стійкий до вибору k (k = 1–21: точність 0,95–0,97), а всі помилки логістичної регресії пов'язані з класом 2. 4. Для подальшої роботи доцільно застосувати той самий протокол до набору Wine Quality [1].

Список використаних джерел

1. Cortez P., Cerdeira A., Almeida F., Matos T., Reis J. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems. 2009. Vol. 47, No. 4. P. 547-553. DOI: 10.1016/j.dss.2009.05.016.

2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York : Springer, 2009. 745 p.

3. Aeberhard S., Forina M. Wine. UCI Machine Learning Repository. 1991. URL: https://archive.ics.uci.edu/dataset/109/wine (дата звернення: 16.10.2026).

4. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825-2830.

УДК 004.85:658.8

АНАЛІЗ ВПЛИВУ ЧИСЛА КЛАСТЕРІВ НА ЯКІСТЬ СЕГМЕНТАЦІЇ СПОЖИВАЧІВ ІНТЕРНЕТ-МАГАЗИНУ

Злата Волохова

Миколаївський національний аграрний університет, м. Миколаїв, Україна

Науковий керівник: О. О. Жебко,
асистент Миколаївський національний аграрний університет

Інтернет-торгівля накопичує великі обсяги транзакційних даних, які дають змогу відійти від єдиної для всіх маркетингової політики до сегментації споживачів за їхньою фактичною поведінкою. Найпоширенішим підходом є RFM-аналіз, у якому кожного клієнта описують давністю останньої

покупки (Recency), частотою покупок (Frequency) та сумою витрат (Monetary) [1]. Поєднання RFM-ознак із кластеризацією методом k-means дає змогу автоматично виділяти групи клієнтів і застосовувати до них різні пропозиції та програми лояльності [2]. Проте метод k-means потребує попереднього задання числа кластерів k, і саме цей гіперпараметр значною мірою визначає якість та практичну цінність сегментації: надто мале k об'єднує різнорідних клієнтів, а надто велике – дробить однорідні групи на штучні сегменти.

Мета роботи – дослідити, як число кластерів впливає на якість сегментації споживачів інтернет-магазину, та обґрунтувати підхід до вибору k. Для цього поставлено завдання: сформувати RFM-ознаки клієнтів; виконати кластеризацію для k від 2 до 10; оцінити якість розбиття за кількома внутрішніми критеріями; порівняти результати й сформулювати рекомендації.

Кожного клієнта описано вектором ознак $x = (R, F, M)$. Алгоритм k-means розбиває множину з n клієнтів на k кластерів $C_1, \ldots, C_k$ так, щоб мінімізувати сумарний внутрішньокластерний квадрат відстаней (1). Для зменшення чутливості до початкового вибору центрів застосовано ініціалізацію k-means++, яка забезпечує оцінку якості розв'язку O(log k) відносно оптимального [3].

$$J(k) = \sum_{j=1..k} \sum_{x \in C_j} \|x - \mu_j\|^2, \qquad (1)$$

де μ_j – центроїд кластера C_j.

Величина $J(k)$ монотонно спадає зі зростанням k, тому сама по собі не дозволяє вибрати число кластерів і використовується в методі ліктя – пошуку точки, після якої спад суттєво сповільнюється. Для об'єктивнішої оцінки застосовано внутрішні критерії. Коефіцієнт силуету [4] для об'єкта i обчислюється за формулою (2), де $a(i)$ – середня відстань до об'єктів власного кластера, $b(i)$ – найменша середня відстань до об'єктів іншого кластера; значення, ближчі до 1, свідчать про компактніші й краще відокремлені кластери, а середній силует може слугувати для вибору k [4].

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}. \qquad (2)$$

Індекс Девіса–Болдіна [5] визначає середню подібність кожного кластера до найближчого сусіднього, тому кращому розбиттю відповідають менші значення. Додатково використано індекс Калінскі–Гарабаша (відношення міжкластерного розкиду до внутрішньокластерного; більші значення кращі) [6]. До інших підходів належить статистика розриву, що порівнює внутрішньокластерний розкид із очікуваним для еталонного розподілу [7]. Розрахунки виконано мовою Python із бібліотекою scikit-learn [8].

Для експерименту сформовано синтетичний набір даних, що імітує RFM-профілі 5000 клієнтів інтернет-магазину з чотирьох заздалегідь заданих груп: активні лояльні (20 %), середні (30 %), нові (25 %) та втрачені клієнти (25 %). Значення R, F, M генеруються логнормальними розподілами з різними медіанами, що відтворює типову правосторонню асиметрію транзакційних даних. Перед кластеризацією ознаки логарифмовано ($\log(1 + x)$) і стандартизовано. Оскільки істинне число груп відоме (чотири), можна перевірити, чи критерії його виявляють, а також порівняти знайдені кластери із заданими за скоригованим індексом Ренда (ARI). Для кожного $k = 2, \ldots, 10$ виконано k-means (ініціалізація k-means++, 10 запусків, фіксоване початкове значення генератора випадкових чисел); результати наведено в табл. 1 і на рис. 1.

Таблиця 1 – Показники якості кластеризації залежно від k

k	J(k)	Силует	Девіс–Болдін	Калінскі–Гарабаш	ARI
2	6627	0,551	0,566	6315	0,273
3	3939	0,487	0,775	7017	0,651
4	**1789**	**0,589**	**0,574**	**12301**	**0,982**
5	1653	0,475	0,991	10084	0,859
6	1530	0,402	1,212	8793	0,797
7	1427	0,318	1,401	7912	0,697

k	J(k)	Силует	Девіс–Болдін	Калінскі–Гарабаш	ARI
8	1326	0,235	1,545	7355	0,589
9	1239	0,238	1,446	6927	0,531
10	1181	0,238	1,386	6491	0,491



Рис. 1. Залежність суми квадратів, середнього силуету та індексу Девіса–Болдіна від k (пунктир – k = 4)

Залежність якості від k не є монотонною. Сума квадратів $J(k)$ різко зменшується при переході від $k = 3$ до $k = 4$ (на 54,6 %), тоді як перехід до $k = 5$ дає лише 7,6 %, тобто «лікоть» спостерігається при $k = 4$. Середній силует досягає максимуму 0,589 при $k = 4$, далі знижується до 0,475 при $k = 5$ і стабілізується на рівні близько 0,24 при $k \geq 8$, що вказує на слабку структуру кластерів [4]. Індекс Калінскі–Гарабаша також максимальний при $k = 4$ (12301). Індекс Девіса–Болдіна має мінімум при $k = 2$ (0,566), проте при $k = 4$ майже не відрізняється (0,574), а далі зростає до 1,54 при $k = 8$; імовірно, грубе розбиття на два макросегменти також є добре відокремленим. Таким чином, три з чотирьох підходів вказують на $k = 4$. Показовим є й збіг із заданою структурою: ARI становить 0,982 при $k = 4$, тоді як при $k = 2$ лише 0,273, а при $k = 10$ – 0,491, тобто як недостатнє, так і надмірне число кластерів спотворює сегментацію.

З практичної точки зору при $k = 4$ кожен кластер має змістовну інтерпретацію. Медіанні профілі в початкових одиницях такі: клієнти з давністю 12 днів, 14 замовленнями та сумою близько 9,6 тис. грн (активні лояльні); 45 днів, 4 замовлення, 2,8 тис. грн (середні); 20 днів, 2 замовлення, 0,7 тис. грн (нові); 187 днів, 2 замовлення, 1,2 тис. грн (втрачені). Для кожної групи можна сформувати окрему стратегію утримання чи реактивації. При $k \geq 6$ виникають сегменти, які розділяють однорідні групи й не мають чіткої бізнес-інтерпретації. Оскільки дані модельні, на реальних даних екстремуми критеріїв виражені слабше, тому вибір k слід підтверджувати також стабільністю кластерів за повторних запусків та змістовністю сегментів.

Висновки. Число кластерів істотно впливає на якість сегментації споживачів: на модельних RFM-даних найкращі значення силуету (0,589), індексу Калінскі–Гарабаша та ARI (0,982) отримано при $k = 4$, що збігається із заданою структурою, тоді як при $k = 8$–10 середній силует знижується до 0,24. Жоден критерій не слід застосовувати ізольовано: мінімум індексу Девіса–Болдіна при $k = 2$ демонструє, що вибір k потрібно ґрунтувати на сукупності методу ліктя, силуету та інших індексів і на бізнес-інтерпретації сегментів. Подальші дослідження передбачають перевірку підходу на реальних транзакційних даних інтернет-магазину та порівняння зі статистикою розриву [7].

Список використаних джерел

1. RFM-аналіз: вчимося сегментувати клієнтів. SendPulse. URL: https://sendpulse.ua/ua/blog/guide-to-building-an-rfm-analysis (дата звернення: 08.10.2026).

2. Serwah A. M. A., Khaw K. W., Cheang S. P. Y., Alnoor A. Customer analytics for online retailers using weighted k-means and RFM analysis. Data Analytics and Applied Mathematics. 2023. Vol. 4, No. 1. P. 1–6. DOI: 10.15282/daam.v4i1.9171. URL: https://journal.ump.edu.my/index.php/daam/article/view/9171 (дата звернення: 08.10.2026).

3. Arthur D., Vassilvitskii S. k-means++: the advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA ’07). Philadelphia : SIAM, 2007. P. 1027–1035. URL: https://doi.org/10.1145/1283383.1283494 (дата звернення: 08.10.2026).

4. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987. Vol. 20. P. 53–65. URL: https://doi.org/10.1016/0377-0427(87)90125-7 (дата звернення: 08.10.2026).

5. Davies D. L., Bouldin D. W. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979. Vol. PAMI-1, No. 2. P. 224–227. URL: https://doi.org/10.1109/TPAMI.1979.4766909 (дата звернення: 08.10.2026).

6. Clustering. Scikit-learn documentation. URL: https://scikit-learn.org/stable/modules/clustering.html (дата звернення: 08.10.2026).

7. Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B. 2001. Vol. 63, No. 2. P. 411–423. URL: https://doi.org/10.1111/1467-9868.00293 (дата звернення: 08.10.2026).

8. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html (дата звернення: 08.10.2026).

УДК 004.932:004.85

ПОРІВНЯННЯ КОМПАКТНОЇ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ ТА LBP-КЛАСИФІКАТОРА ЗА ОБСЯГОМ НАВЧАЛЬНОЇ ВИБІРКИ ПРИ КЛАСИФІКАЦІЇ ТЕКСТУР МАТЕРІАЛІВ

Владислав Макеєв, Олександр Жебко

Миколаївський національний аграрний університет, м. Миколаїв, Україна

Автоматичне розпізнавання матеріалів за зображеннями їхніх текстур застосовують у контролі якості, сортуванні сировини та агропромисловому виробництві. Глибокі згорткові нейронні мережі (CNN) стали основним інструментом такої класифікації [2, 3], однак вони потребують підбору навчальних даних, тоді як класичні дескриптори, як-от локальні бінарні шаблони (LBP) [1], мають вбудовану інваріантність до яскравості та повороту. На практиці розмічених зразків часто мало, тому важливо знати, з якого обсягу вибірки компактна CNN починає відповідати простому базовому методу.

Текстура задається статистикою локальних змін яскравості, тому її опис залежить від масштабу й умов освітлення. Дескриптор LBP порівнює яскравість кожного пікселя з яскравістю сусідів на колі заданого радіуса та кодує результат бінарним числом; гістограма таких кодів у фрагменті слугує вектором ознак [1]. Згорткові мережі натомість навчаються витягувати ознаки безпосередньо з пікселів, а їхні перші шари можна розглядати як банк фільтрів, параметри якого підбираються під задачу [2]. Для оцінювання таких підходів зазвичай використовують великі колекції, зокрема набір DTD із 5640 зображень у 47 категоріях [3]; водночас у прикладних задачах власних розмічених зображень зазвичай значно менше.

Мета дослідження – кількісно порівняти точність компактної CNN і класифікатора LBP+SVM залежно від кількості навчальних фрагментів текстури та перевірити стійкість обох моделей до адитивного шуму.

Використано чотири реальні напівтонові зображення 512×512 пікселів із відкритої бібліотеки scikit-image [4]: цегляна кладка, трава, гравій і поверхня Місяця. Фрагменти розміром 64×64 пікселі стандартизували (нульове середнє, одиничне стандартне відхилення) і квантували до 8 біт. Щоб уникнути витоку даних, навчальні фрагменти випадково добирали зі стовпців 0–319 кожного зображення, а тестові – зі стовпців 352–511 з кроком 16 пікселів (відступ 32 пікселі), що дало 812 тестових фрагментів (по 203 на клас). Кількість навчальних фрагментів на клас n змінювали від 2 до 256; для кожного n виконано 10 незалежних запусків з різними вибірками та ініціалізаціями. CNN